IBM Docket No.: AUS920030892US1
Express Mail No. EL977166125US

Attorney Docket No.: IBM.4021.PAT

UNITED STATES PATENT APPLICATION

for

GLOBAL MANAGEMENT OF LOCAL LINK POWER CONSUMPTION

INVENTOR:

Juan-Antonio Carballo

Certification Under 37 CFR 1.10

EXPRESS MAIL CERTIFICATE OF MAILING

"Express Mail" Mailing Label Number EL977166125US Date of Deposit Dec. 22, 2003

I hereby certify that this New Application and the documents referred to as enclosed therein are being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service on the date indicated above and addressed to Mail Stop Patent Application, Commissioner for Patents, P. O. Box 1450, Alexandria, VA 22313-1450.

Jeffrey S. Schubert
(Typed or printed name of person mailing paper or fee)

(Signature of person mailing paper or fee)

GLOBAL MANAGEMENT OF LOCAL LINK POWER CONSUMPTION

FIELD OF INVENTION

[0001] The present invention is in the field of communications channels, or links. More particularly, the present invention relates to methods and arrangements for power reduction in links, such as transmitters and receivers, based upon global decisions such as the data transmission frequency, communications media, and traffic type associated with the links.

BACKGROUND

[0002] Communication systems typically include logic and hardware to transmit data from an origin to a destination. In particular, communication systems have routing or switching logic to make high-level decisions that select ports, routes, and media for transmitting the data. Communication systems also include links, each having a transmitter, a medium, and a receiver, to transmit the data in response to those high-level decisions.

[0003] The origin clocks the data originally. Then, each intermediate link, or more specifically, the link's transmitter typically clocks the data and transmits to the link's receiver or the destination.

[0004] Devices such as routers typically access a network identification (NETID) for the data transmission to determine the destination and calculate the route to the destination through intermediate links based upon a routing protocol and a routing table that includes information about the communication system's topology. The routing protocol dynamically determines routing for the data transmission, taking into consideration changing conditions of the communication network such as unavailable links. Routing tables, for instance, may associate links with ports, or port numbers, through which the data transmission should be routed.

[0005] Upon determining the port for the data transmission, the transmission is routed through that port to the destination or another, intermediate destination. Some of the more complex routers, such as routers for super computers, may also select a medium through which the transmitter and receiver will transmit the data.

[0006] The transmitters and receivers may consume more power depending upon the data transmission and the media through which the data transmission is routed. In particular, data transmissions at higher data frequencies, with difficult data traffic types or patterns, via long media, and/or via lossy media, require amplifiers and complex, mixed-signal circuitry. The amplifiers and complex, mixed-signal circuitry improve or maximize the sampling window for bits of data in the data transmission to maintain an acceptable bit error rate (BER), i.e., the number of misinterpreted bit values for the data transmission.

[0007] Higher data frequencies require internal circuits of transmitters and receivers to operate at high clock frequencies and, thus, high voltage levels, to sample and re-transmit the data in each intermediate link. Further, when the clock frequencies of a transmitter and receiver pair have differences in phase that change over time, often referred to as spread spectrum signaling, the receiver may include a clock and data recovery (CDR) loop with second and, possibly, third order frequency tracking circuits running at high internal frequencies.

[0008] Similarly, demanding traffic types, which include patterns that do not often switch between logical ones and zeros or that switch between ones and zeros in irregular or sporadic patterns, require complex internal circuits of transmitters and receivers that may operate at high clock frequencies to capture the relatively few transitions. The phase of the sampling clock is adjusted based upon the phase of the data signal as determined from those relatively few transitions.

[0009] With regards to long and/or lossy media, the amplitude of the data transmission may attenuate in a frequency-dependent manner. Amplification and pre-emphasis by the transmitter as well as amplification and equalization by the receiver accentuate certain

frequencies to increase the sampling window. Other circuitry such an internal loop filters may be more complex when the media is long and/or lossy.

[0010] The amplification and complex, mixed-signal circuitry, however, significantly increase the overall power consumption for the communication system. For example, serial links within a large interconnect system such as a super computer may consume 20 to 37% of total power consumption.

[0011] Further, the amplifiers and complex, mixed-signal circuitry continue to operate at full power even when such circuitry is unnecessary. For example, a high-level decision may make a link inactive or switch the media for the link from a long medium that requires the complex, mixed-signal circuitry, to a shorter medium that does not require such circuitry.

[0012] Thus, there is a need for methods and arrangements for power reduction in link circuits such as transmitters and receivers based upon global, or high-level, decisions such as the activity, data transmission frequency, communications media, and traffic type associated with links.

SUMMARY OF THE INVENTION

[0013] The problems identified above are in large part addressed by methods and arrangements for power reduction in links, such as transmitters and receivers, based upon global decisions such as the activity, data transmission frequency, communications media, and traffic type associated with links. One embodiment provides a local link for reducing power consumption. The variable power link contemplates a link circuit to process data having multiple different data transmission characteristics, the link circuit being configurable to operate in multiple power modes, wherein at least two of the multiple power modes are associated with respective data transmission characteristics; and a local controller to receive activity assignments for the variable power link, wherein the activity assignments are related to data transmission characteristics, and to configure the link circuit to operate in one of the multiple power modes in respective response to a received activity assignment.

[0014] Another embodiment provides an apparatus for reducing power consumption by a link. The apparatus contemplates a port utilization manager to track an availability of a port; forwarding logic to associate the port with a destination; and a global controller coupled with the forwarding logic to determine an activity for a link based upon an association between the link and the port and the availability of the port, the activity being related to a data transmission characteristic for data to transmit via a channel of the link, and to transmit a control signal to a local controller, wherein the control signal indicates a power mode for circuitry associated with the link and the data is associated with the destination.

[0015] Another embodiment provides a method for reducing power consumption by a link. The method contemplates determining an activity for the link based upon forwarding logic, the activity being related to a characteristic for a data transmission via a channel of the link; associating the activity with a power mode for the link, wherein the power mode is related to the characteristic; and configuring circuitry associated with the link to operate in the power mode to process the data transmission.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] Other objects and advantages of the invention will become apparent upon reading the following detailed description and upon reference to the accompanying drawings in which, like references may indicate similar elements:

FIG 1 depicts an embodiment of a system including a router and a hub to transmit data from one processor card to another;

FIG 2 depicts an embodiment of a network component such as the router in FIG 1 for reducing power consumption by a link;

FIG 3 depicts an embodiment of a flow chart for a global controller for reducing power consumption by a link; and

FIG 4 depicts an embodiment of a flow chart for a local controller for reducing power consumption by a link.

DETAILED DESCRIPTION OF EMBODIMENTS

[0017] The following is a detailed description of example embodiments of the invention depicted in the accompanying drawings. The example embodiments are in such detail as to clearly communicate the invention. However, the amount of detail offered is not intended to limit the anticipated variations of embodiments, but on the contrary, the intention is to cover all modifications, equivalents, and alternatives falling within the spirit and scope of the present invention as defined by the appended claims. The detailed descriptions below are designed to make such embodiments obvious to a person of ordinary skill in the art.

[0018] Generally speaking, methods and arrangements for power reduction in links, such as transmitters and receivers, based upon global decisions such as the data transmission frequency, communications media, and traffic types associated with links, are contemplated. In particular, embodiments take advantage of high-level decisions to reconfigure internal circuits of links to reduce power consumption. At the global or system level, a decision determines the links that are active (i.e., turned on or off), the data frequency at which the links are operating, and the media through which the links transmit the data. These decisions are then communicated to the links.

[0019] At the local level, the links receive the decisions and each transmitter and receiver pair reconfigures internal circuitry automatically to minimize power based upon the decisions. In some embodiments, the links may receive the decisions in the form of power modes. For example, each link may receive a control signal indicating that the link should be in a turned off mode, a low power mode, and/or a standard power mode. In further embodiments, the links may receive settings such as an on/off setting, a data frequency setting, and a traffic/media setting. In such embodiments, the combination of the settings may indicate a power mode so, based upon

the settings, the link may selectively adjust or modify the operation of the internal circuits of the transmitter and receiver.

[0020] While specific embodiments will be described below with reference to particular circuit configurations, power modes, and combinations of settings, those of skill in the art will realize that embodiments of the present invention may advantageously be implemented with other, substantially equivalent circuit configurations, power modes, and settings. In particular, the settings may be directly associated with a circuit of a transmitter and/or receiver and may, in some embodiments, indicate particular modifications to the circuit such as a change in the clock frequency for the circuit or a change in the voltage of the high voltage source, Vdd.

[0021] Turning now to the drawings, FIG 1 depicts an embodiment of a system 100 including a router 110 and a hub 140 to transmit data from a processor card 105 to a processor card 170. For example, processor card may transmit data 107 to processor card 170 via router 110 and hub 140. Certain details such as data buffers are not shown explicitly for simplicity.

[0022] Router 110 may determine links through which the data 107 will be transmitted and adjust the operation of link circuits of the links to correlate power consumption by the links with characteristics of the data transmission. Router 110 may include routing table 112, port utilization 114, global link control 116, read port 120 and write port 130.

[0023] Routing table 110 may include a database that contains the current network topology, to direct packets of a data transmission out the appropriate port. Routing 110 may determine the appropriate path onto which data 107 should be forwarded from routing table 112 based upon a routing protocol. The routing protocol may also allow the network to dynamically adjust to changing conditions by describing how routers share updated information about the topology. For instance, routing table may indicate a route from processor card 105 to processor card 170 via read port 120, write port 130, read port 150, and write port 160.

[0024] Port utilization manager 114 may track availability and usage of ports such as ports 120, 130, 150 and 160. In the present embodiment, port utilization manager 114 is a global logic for tracking port utilization. In further embodiments, port utilization manager 114 may include logic incorporated into more than one switching chips of router 110 or a combination of global and switching chip level logic. For instance, router 110 may include, e.g., several switching chips, each having 100 ports and a port utilization manager that communicates with global link control. In other embodiments, port utilization manager 114 may include a signal received from logic exterior to router 110.

[0025] Based upon the availability of links and the routing information, global link control 116 may determine that ports 120, 130, 150, and 160 will transmit data 107 from processor card 105 to processor card 170. Global link control 116 may then gather information about data 107 to configure link circuits 124, 134, 154, and 164. In particular, global link control 116 may receive data from processor card 105 that describes data 107. For instance, global link control 116 may determine the type of encoding used to encode data 107 and based upon that encoding, determine the type of data traffic associated with transmitting data 107. For example, a real time video or audio encoding may include long strings of logical ones or zeroes that act as a filler for a video stream to describe the passage of time. The long strings of logical ones or zeroes are a difficult traffic condition because of the small number of transitions per unit time, providing clock and data recovery (CDR) loops little information for maintaining the phase of a sampling clock utilized to sample values for each bit in the data stream. Further types of encoding may produce irregular or sporadic patterns of bit values that are also difficult for a receiver to decipher.

[0026] In additional to determining the traffic type, global link controller 116 may also determine, e.g., a data frequency at which to transmit data 107. For instance, global link control 116 may determine a rate at which processor card 105 can transmit data 107 as well as the limitations on data frequency, or bandwidth, throughout the links between processor card 105 and processor card 170.

[0027] Once global link control 116 determines data transmission characteristics such as the traffic type, the data frequency, and the route for data 107, global link control 116 may determine a power mode for ports 120, 130, 150, and 160 to correlate power consumption with the characteristics or constraints of transmission of data 107. More specifically, when the traffic type is difficult such as the long strings of logical ones and zeroes or the data frequency is high, more complex logic and circuitry of link circuits 124, 134, 154, and 164 may be powered with higher voltages and clocked with higher clock frequencies to handle the data transmission. In such situations, a standard power mode may be selected and an indication of the standard power mode may be transmitted in a control signal to ports 120, 130, 150, and 160.

[0028] On the other hand, when the traffic type is not difficult, or is simple, and the data frequency is not sufficiently high to justify the use of the more complex logic and circuitry of link circuits 124, 134, 154, and 164, global link control 116 may transmit a control signal to ports 120, 130, 150, and 160 to indicate a low power mode. Further, global link control 116 may determine that other links (not shown) may not be needed to transmit data 107 so they may be turned off. In some of those situations, global link control 116 may also transmit a control signal to turn off and/or reduce power to circuitry of link circuits 124, 134, 154, and 164.

[0029] Ports such as read ports 120 and 150, and write ports 130 and 160 may include receivers and transmitters designed to respond to control signals from global link control 116 by configuring and/or re-configuring link circuits based upon the power mode indicated by the control signals. For example, local link control 122 may receive a control signal from global link control 116 to configure link circuit 124 to a low power mode and, in response, local link control 122 may, for instance, reduce the amplification of data 107.

[0030] Similarly, local link control 122 may receive a control signal from global link control 116 to re-configure link circuit 124 to a standard power mode. Re-configuring link circuit 124 to the standard power mode may facilitate transmission of data 107 at a higher data frequency.

[0031] After port 120 is configured to handle data 107, data 107 may be transmitted from processor card 105 to read port 120. Then, data 107 may be transmitted through ports 130, 150, and 160 to processor card 170. Advantageously, the ports 120, 130, 150, and 160 consume power at a rate based upon the routing decision of global link control 116 that corresponds to the difficulty in transmitting data 107 through system 100.

[0032] FIG 2 depicts an embodiment of a link coupled with a global decision device 210 such as router 110 in FIG 1. Link 200 includes a global decision device 210 and a link 219. For example, global decision device 210 makes a routing decision that link 219 is to transmit data at three Gbps instead of ten Gbps. The routing decision involves changing the data frequency of link 219 from ten Gbps to three Gbps and link 219 includes transmitter 220 and receiver 250. Local link control 222 of transmitter 220 receives the decision 216 and, in response, turns down the gain of an analog amplifier for driver 228 and turns off pre-emphasis circuit 226. Similarly, local link control 252 of receiver 250 receives the decision 218 and, in response, turns down the analog, receiver amplifier 254 and turns off gain and equalization circuit 256. Advantageously, based upon the high-level, routing decision of global decision device 210, power consumption of link 219 is reduced.

[0033] Global decision device 210 may make a global decision regarding an activity assignment for link 219 based upon information about port utilization 205 and forwarding logic 212, and transmit a control signal 216 and 218 to link 219 to indicate the activity assignment for the link. More specifically, global decision device 210 be part of a switch or router and comprise global link control 214 to determine data transmission characteristics such as the data frequency associated with link 219, the data traffic for link 219, and the medium through which data transmission 240 is transmitted. For example, global link control 214 may determine whether to turn off link 219 based upon destinations associated with incoming data transmissions and network topology information that associates ports between link 219 and the destination via forwarding logic 212. Global link control 214 may select a data frequency for link 219 based upon a data frequency of an incoming data transmission. And, global link control 214 may read

packet headers of the incoming data transmission to determine whether a traffic pattern is a difficult pattern or a simple pattern.

[0034] Upon determining whether to turn link 219 off, the data frequency for link 219, and whether the traffic pattern is simple or difficult, these operating parameters may be transmitted in a control signal 216 directly to transmitter 220 and in a control signal 218 directly to receiver 250. In other embodiments, the operation parameters may be transmitted to transmitter 220 and transmitter 220 may communicate the operation parameters, or an indication thereof, to receiver 250, or vice versa.

[0035] The operating parameters resulting from the routing decision may be associated with a power mode for circuits of link 219 by interpretation logic 223 of transmitter 220 and interpretation logic 253 of receiver 250. More specifically, local link control 222 may receive the operation parameters and utilize interpretation logic 223 to translate the operation parameters into a power mode for serialization circuit 224, pre-emphasis circuit 226, and driver 228. For instance, when the operation parameters indicate that link 219 is to be turned off, interpretation logic 223 may indicate that one or more circuits of serialization circuit 224, pre-emphasis circuit 226, and driver 228 should be turned off, declocked, and/or operated at a minimum voltage and frequency.

[0036] On the other hand, when the operation parameters indicate that link 219 is to be turned on and to operate at a high frequency or transmit a difficult traffic pattern of data, interpretation logic 223 may indicate that pre-emphasis circuit 226 is turned on and operating at a high complexity level, and driver 228 is operating at a high gain. Operating pre-emphasis circuit 226 at a high complexity level and driver 228 at a high gain may involve increasing one or more clock frequencies associated with pre-emphasis circuit 226 and driver 228, and increasing the high voltage source(s) for the circuits in conjunction with increasing the frequency.

[0037] Similarly, receiver 250 may receive the operating parameters either directly from global decision device 210 via control signal 218, or from transmitter 220. In response to receiving operation parameters indicative of a power mode for receiver 250, or a circuit of receiver 250, interpretation logic 253 of local link control 252 may determine a configuration for a receiver amplifier 254, a gain and equalization circuit 256, and a clock and data recovery (CDR) loop 258. For example, receiver 250 may receive operating parameters describing link 219 as being turned on with a low data frequency and a simple traffic pattern. In response, receiver amplifier 254 and CDR loop 258 are reduced to a minimum gain and gain and equalization circuit 256 is turned off or reduced to a minimum functionality. In some embodiments, when CDR loop 258 is turned off, a substitute CDR loop having minimum functionality is enabled for receiver 250.

[0038] Some time after deciding that link 219 should be configured for a low data frequency and a simple traffic type, global decision device 210 may decide to change the activity of link 219. For instance, the data frequency for data transmissions between transmitter 220 and receiver 250 may vary between three Gigabits per second (Gbps) and six Gbps and when the data frequency of an incoming data transmission is at six Gbps, global decision device 210 may determine that link 219 should operate at a data frequency of six Gbps to transmit the data 230 from the incoming data transmission to receiver 250 via data transmission 240. Global link control 214 generates a control signal 216 for transmitter 220 and a control signal 218 for receiver 250. The medium type may be long or particularly lossy. And the traffic type may complicate clocking the data transmission, e.g., the data traffic may be very active, often switching between logical ones and logical zeroes at varying frequencies. Thus, global link control 214 may determine operating parameters that indicate a high data frequency and a difficult traffic type. In response, local link control 222 may turn on pre-emphasis circuit 226 and raise the frequencies and high voltage source for pre-emphasis circuit 226 to maximum ratings. Similarly, local link control 222 may increase the bias for driver 228 to a maximum.

[0039] Receiver 250 may receive the operating parameters via control signal 218. In response to control signal 218, local link control 252 may increase the bias for receiver amplifier

254 to a maximum, increase the complexity, high voltage source, and/or frequencies for gain and equalization circuit 256, and implement a circuit for CDR loop 258 having second and third order frequency tracking.

[0040] Upon adjusting the activity of link 219, serialization circuit 224 serializes data 230, clocking the data at six Gbps, and, in many embodiments, pre-amplifies the serialized data based upon input specifications for pre-emphasis circuit 226. In many embodiments, for instance, serialization circuit 224 includes a low frequency clock source having low jitter (to maximize the sampling window for the data) and a multiple phase output. The rising and/or falling edges of the multiple phases are utilized to clock parallel inputs of data 130 into a single, six Gbps data stream.

[0041] Pre-emphasis circuit 226 may utilize a finite impulse response (FIR) equalizing filter to cancel or at least reduce frequency-dependent attenuation such as attenuation caused by the skin-effect resistance of copper wire when copper wire is the medium through which data transmission 240 is transmitted. Pre-emphasis circuit 226 accentuates the high frequency components of the data signal to at least partially alleviate the effects of inter-symbol interference (ISI).

[0042] Then, driver 228 drives data transmission 240 across a medium such as a copper wire or an optical fiber. In some embodiments, for example, driver 228 may include a Fibre Channel driver and data transmission 240 may be transmitted through a channel of a fiber optic cable.

[0043] Receiver 254 receives data transmission 240 from driver 228 and pre-amplifies data transmission 240 for gain and equalization circuit 256. Gain and equalization circuit 256 amplifies data transmission 240 and accentuates the high frequency components to attempt to increase the sampling window for the data. Then, CDR loop 258 samples the data from the data signal, compares the phase of the sampling clock to the phase of the data transmission 240 and adjusts the sampling clock accordingly. When second order and third order frequency tracking

circuits are included in CDR loop 258, second and third order corrections are made to adjustments of the sampling clock phase. For example, initial samples from the data transmission indicate instantaneous, high frequency changes to the phase of the data transmission. Second order and third order frequency tracking circuits observe and correct for lower frequency changes in the phase of data transmission 240. Once CDR loop 258 samples the data transmission 240, the determined values of the bits are output as data 260.

[0044] Referring now to FIG 3, there is shown an example of a flow chart 300 for a global controller such as router 110 of FIG 1. The global controller may determine activity assignments for each link and the activity assignments may be communicated to the local controller of the link in the form of one or more power modes or one or more settings that are indicative of power modes for link circuits such as circuits for pre-emphasis, amplification, equalization, and CDR. Flow chart 300 begins with turning off unnecessary links (element 310). More specifically, a high-level decision device determines which links to turn off based upon forwarding logic such as a routing table and the current utilization of ports associated with the links.

[0045] After the unnecessary links are turned off, the high-level decision device assigns destination nodes to active links to utilize the links for different, incoming data transmissions based upon forwarding logic (element 320). Selecting destination nodes for the active links may involve assigning particular incoming data transmissions to ports associated with particular active links. In further embodiments, assigning destination nodes to a link involves selecting a medium through which the link will transmit data. Selecting the medium may be more prevalent, for example, in optical routers.

[0046] Once the destination nodes are assigned to ports of active links, characteristics of a data transmission associated with the ports of active links are collected to determine an activity assignment for the corresponding active link (element 330). The activity assignment may describe, for instance, the data frequency for the data transmission, and the media type and traffic

type of the data transmission. In some embodiments, the activity assignment may be represented by a power mode.

[0047] The activity assignment is then transmitted to the link and if there are more links for which the global decision device determines an activity assignment (element 345), characteristics associated with the data transmission for those links are determined (element 330) and transmitted (element 340).

[0048] Referring now to FIG 4, there is shown an example of a flow chart 400 for local controller such as transmitter 220 or receiver 250 as shown in FIG 2. The local controller may include a local link control such as local link control 222 or local link control 252 in FIG 2. For example, the local controller may receive an activity assignment from a high–level decision device such as a router that determines what data will be transmitted via a link associated with the local controller, and what medium will be used to transmit that data to a destination.

[0049] Flow chart 400 begins with receiving an activity assignment (element 410) for a link. The activity assignment may indicate a data frequency, a media type and/or a traffic type for a data transmission to be processed by one or more link circuits associated with the link.

[0050] Based upon the activity assignment, the power modes of each of the one or more link circuits may be determined (element 420). For instance, when the activity assignment includes a data frequency and the data frequency is a relatively high frequency, link circuits may be configured to operate in a high power consumption state to process the high data frequency. Similarly, if the activity assignment includes an indication that the traffic type is difficult, link circuits may be configured to operate in a high power consumption state to amplify, pre-emphasize and/or equalize the data transmission.

[0051] Upon determining the power mode for the link circuit, the link circuit is configured based upon the power mode to adjust power consumption (element 430). Whether the power consumption of the link circuit is high to accommodate high data frequencies or

difficult traffic types, the link circuit is dynamically configured to, advantageously, consume an amount of power related to the complexity of the data transmission.

[0052] Then, if the link includes additional link circuits, the additional link circuits are configured based upon the activity assignment communicated by the high-level decision device (element 440).

[0053] It will be apparent to those skilled in the art having the benefit of this disclosure that the present invention contemplates methods and arrangements for power reduction in links, such as transmitters and receivers, based upon global decisions such as the activity, data transmission frequency, communications media, and traffic type associated with links. It is understood that the form of the invention shown and described in the detailed description and the drawings are to be taken merely as examples. It is intended that the following claims be interpreted broadly to embrace all the variations of the example embodiments disclosed.